APPLICATION OF


FRANK M. ZIZZAMIA
A Citizen of the United States
Residing at
81 Wheeler Road
Avon, CT 06001

CHENG-SHENG PETER WU
A Citizen of the United States
Residing at
1720 Highland Oaks
Arcadia, CA 91006

RAYMOND E. STUKEL
A Citizen of the United States
Residing at
728 West Jackson Blvd., Apt. 426
Chicago, IL 60661

HRISANTHI ADAMOPOULOS
A Citizen of the United States
Residing at
78 Village Drive, Apt. 403
Wethersfield, CT 06109


FOR LETTERS PATENT OF THE UNITED STATES

FOR IMPROVEMENTS IN

**METHOD AND SYSTEM FOR DETERMINING THE IMPORTANCE OF INDIVIDUAL VARIABLES IN A STATISTICAL MODEL**


Randy Lipsitz, Esq.
Registration No. 29,189
John C. Garces, Esq.
Registration No. 40,616
Attorneys for Applicants
KRAMER LEVIN NAFTALIS & FRANKEL LLP
919 Third Avenue
New York, New York 10022
Telephone No. (212) 715-9100


Docket No. 098056/00120

## METHOD AND SYSTEM FOR DETERMINING THE IMPORTANCE OF INDIVIDUAL VARIABLES IN A STATISTICAL MODEL

BACKGROUND OF THE INVENTION

The present invention is directed to a method and system for evaluating the results of a predictive statistical scoring model and more particularly to a system and method that determines the contribution of each of the variables that comprise the predictive scoring model to the overall score generated by the model.

Insurance companies provide coverage for many different types of exposures. These include several major lines of coverage, e.g., property, general liability, automobile, and workers compensation, which include many more types of sub-coverage. There are also many other types of specialty coverages. Each of these types of coverage must be priced, i.e., a premium selected that accurately reflects the risk associated with issuing the coverage or policy. Ideally, an insurance company would price the coverage based on a policyholder's actual future losses. Since a policyholder's future losses can only be estimated, an element of uncertainty or imprecision is introduced in the pricing of a particular type of coverage such that certain policies are priced correctly, while others are under-priced or over-priced.

In the insurance industry, a common approach to pricing a policy is to develop or create complex scoring models or algorithms that generate a value or score that is indicative of the expected future losses associated with a policy. The predictive scoring models are used to price coverage for a new policyholder or an existing policyholder. As is known, multivariate analysis techniques such as linear regression, nonlinear regression, and neural networks are commonly used to model insurance policy profitability. A typical insurance profitability application will contain many predictive variables. A profitability application may be comprised of thirty to sixty different variables contributing to the analysis.

The potential target variables in such models can include frequency (number of claims per premium or exposure), severity (average loss amount per claim), or loss ratio (loss divided by premium). The algorithm or formula will directly predict the target variable in the model. The scoring formula contains a series of parameters that are mathematically combined with the predictive variables for a given policyholder to determine the predicted profitability or final score. Various mathematical functions and operations can be used to produce the final score. For example, linear regression uses addition and subtraction operations, while neural networks involve the use of functions or options that are more complex such as sigmoid or hyperbolic functions and exponential operations.

In creating the predictive model, often the predictive variables that comprise the scoring formula or algorithm are selected from a larger pool of variables for their statistical significance to the likelihood that a particular policyholder will have future losses. Once selected from the larger pool of variables, each of the variables in this subset of variables is assigned a weight in the scoring formula or algorithm based on complex statistical and actuarial transformations. The result is a scoring model that may be used by insurers to determine in a more precise manner the risk associated with a particular policyholder. This risk is represented as a score that is the result of the algorithm or model. Based on this score, an insurer can price the particular coverage or decline coverage, as appropriate.

As noted, the problem of how to adequately price insurance coverage is challenging, often requiring the application of complex and highly technical actuarial transformations. These technical difficulties with pricing coverages are compounded by real world marketplace pressures such as the need to maintain an "ease-of-business-use" process with policyholders and insurers, and the underpricing of coverages by competitors attempting

to buy market share. Notwithstanding the recognized value of these pricing models and their simplicity of use, known models provide insurers with little information as to why a particular policyholder received his or her score. Consequently, insurers are unable to advise policyholders with any precision as to the reason a policyholder has been quoted a high premium, a low premium, or why, in some instances, coverage has been denied. This leaves both insurers and policyholders alike with a feeling of frustration and almost helpless reliance on the model that is used to determine pricing.

While predictive scoring models are available in the insurance industry to assist insurers in pricing insurance coverage, there is still a need for a method and system to that overcomes the foregoing shortcomings in the prior art. Accordingly, there exists a need for a system and method to interpret the results of any scoring model used in the insurance industry to price coverage. Indeed, the system and method may be used to interpret the results of any complex formula. There is especially a need for a system and a method that allow an insurer to determine and rank the contribution of each of the individual predictive variables to the overall score generated by the scoring model. In this manner, insurers and policyholders alike may know with certainty the factors or variables that most influenced the premium paid or price of an insurance policy.



## SUMMARY OF THE INVENTION

It is an object of the present invention to address and overcome the deficiencies of the prior art by providing a system and a method for interpreting the results of a scoring model used to price insurance coverage.

It is another object of the invention to provide a system and a method that determine the significance or contribution of each predictive variable to the score generated by such a scoring model.

It is another object of the invention to provide a system and a method that permit insurers to rank the variables according to their significance or contribution to such overall score.

It is still another object of the present invention to provide a system and method that allow insurers to utilize the rank information to inform potential or existing policyholders of those variables that most influenced or affected the pricing.

Accordingly, in one aspect of the invention a method is provided of evaluating the scoring formula or algorithm to determine the contribution of each of the individual predictive variables to the overall score generated by the scoring model. For example, in the commercial auto industry, sophisticated scoring models are created for predicting the profitability of issuing a particular policy based on variables that have been determined to be predictive of profitability. These predictive variables may include the age of the vehicle owner, total number of drivers, speeding violations and the like. In a scoring algorithm having over a dozen variables, the analysis or their individual contributions to the overall score would be very difficult without the present invention.

In another aspect of the present invention, in a system that employs a statistical model comprised of a scoring formula having a plurality of predictive variables for generating a score that is representative of a risk associated with an insurance policyholder and for pricing a particular coverage based on the score, a method is provided of quantifying the contribution of each of the plurality of predictive variables to the score generated by the model including the steps of populating a database associated with the system with a mean value and standard deviation value for each of the plurality of variables, calculating a slope value for each of the plurality of variables, calculating a deviance value based on the slope and standard deviation for each of the plurality of variables, and multiplying the deviance

value and slope value for each of the plurality of variables to quantify the contribution of each of the plurality of variables to the score. This quantified contribution may then be used to rank the variables by importance to the overall score.

Additional objects, features and advantages of the invention appear from the following detailed disclosure.

The present invention accordingly comprises the various steps and the relation of one or more of such steps with respect to each of the others, and the product which embodies features of construction, combinations of elements, and arrangement of parts which are adapted to effect such steps, all as exemplified in the following detailed disclosure, and the scope of the invention will be indicated in the claims.

## BRIEF DESCRIPTION OF THE DRAWINGS

For a fuller understanding of the invention, reference is made to the following description, taken in connection with the accompanying drawings, in which:

Fig. 1 illustrates a system that may be used to interpret and rank the predictive variables according to an exemplary embodiment of the present invention;

Fig. 2 is a flow diagram depicting the steps carried out in interpreting the contribution of each of the predictive external variables in a scoring model according to an exemplary embodiment of the present invention;

Fig. 3 specifies the description of the variables in an example illustrating the application of the method of the present invention to an exemplary scoring formula;

Fig. 4 specifies assumptions made regarding the variables in the exemplary scoring formula; and

Fig. 5 specifies the values for the variables used in the exemplary scoring formula, the application of the method of the present invention and the results thereof.

## DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS

The invention described herein creates an explanatory system and method to quantitatively interpret the contribution or significance of any particular variable to a policyholder's profitability score (hereinafter the "Importance"). The methodology of the present invention considers both a) the overall impact of a variable to the scoring model as well as b) the particular value of each variable in determining its Importance to the final score.

As is known, scoring models are developed and used by the insurance industry (as well as other industries) to set an ideal price for a coverage. Many off-the -shelf statistical programs and applications are known to assist developers in creating the scoring models. Once created relatively standard or common computer hardware may be used to store and run the scoring model. Fig. 1 illustrates an exemplary system 10 that may be employed to implement a scoring model and calculate the Importance of individual predictive variables according to an exemplary embodiment of the present invention. Referring to Fig. 1, the system includes a database 20 for storing the values for each of the variables in the scoring formula, a processor 30 for calculating the target variable in the scoring algorithm as well as the values associated with the present invention, monitor 40 and input/output means 50 (i.e., keyboard and mouse). Alternatively, the system 10 may housed on a stand alone personal computer having a processor, storage means, monitor and input/output means.

Referring to Fig. 2, the steps of a method according to an exemplary embodiment of the present invention are shown generally as 100. The method assumes a model has been generated utilizing one of many statistical and actuarial techniques briefly

discussed herein and known in the art. The model is typically a scoring formula or algorithm comprised of a plurality of weighted variables. The database 20 is populated with values for the variables that define the scoring model. These values in the database are used by the scoring model to generate the profitability score. It should be noted that some of the values might be supplied as a separate input from an external source or database.

Similarly, in step 101, the database 20 or a different database is populated with values for the population mean and standard deviation for each of the predictive variables. These values will be used in calculating the Importance as will be described. Next, in step 102, the slope for each predictive variable in the scoring model is determined. As discussed below, this may be simply done in a scoring mode or require a separate calculation. In step 103, a deviance is calculated. After the deviance is calculated, in step 104, the Importance is calculated for each variable by multiplying the slope by the deviance. The variables are then ranked by Importance in step 105. The higher the value the more important the variable was toward the overall profitability score.

Steps 102 through 104 are now explained in more detail:

<u>Step 102</u>

The first criterion in determining the most important variables for a particular score is the impact that each variable contributes to the overall scoring formula. Mathematically, such impact is given by the slope of the scoring function with respect to the variable being analyzed. To calculate the slope, the first derivative of the formula with respect to the variable is generated. For a nonlinear profitability formula such as a neural network formula or a nonlinear regression formula, the slope may be different from one data point (i.e. policyholder) to the next. Therefore, the average of the slope across all of the data points is used as the first criteria to measure Importance.

Since the first derivative can be either positive or negative for each data point and since the impact should be treated equally regardless of the sign of the slope, it is necessary to calculate the average of the first derivative and then take the absolute value of the average. In summary, the first criteria in determining the most important variables can be represented as follows: Slope of Predictive Variable $x_i = \left| avg(\frac{\partial F(X)}{\partial x_i}) \right|$ (where *F(X)* is the scoring function which depends on a number of predictive variables, $x_i$, i=1,2,3...n.b).

This technique is also directly applicable to the linear regression model results. However, in a linear regression model, the slope of a variable is constant (same sign and same value) across all of the data points and therefore the average is simply equal to the value of the slope at any particular point.

<u>Step 103</u>

Although the slope impact of a predictive variable as determined in Step 102 is applied to every data point, it is expected that the Importance of any particular variable will be different from one data point to another. Therefore, the overall Importance of a variable should include a measure of its value for each specific policyholder as well as the overall average value determined in Step 102. For example, if the value of a variable deviates "significantly" from the general population mean for a given policyholder, the conclusion might be that the variable played a significant role in determining why that policy received its particular score. On the other hand, if the value of a particular variable for a chosen policy is close to the overall population mean, it should not be judged to have an influential impact on the score, even if the average value of the variable impact (from Step 102) is large, because its value for that policy is similar to the majority of the population.

Therefore, the second criterion in measuring Importance, Deviance, is a measure of how similar or dissimilar a variable is relative to the population mean. Deviance may be calculated using the following formula:

$$\text{Deviance of } x_i = \frac{(x_i - \mu_i)}{\sigma_i}$$

where $\mu_i$ is the mean for $x_i$ and $\sigma_i$ is the standard deviation for predicitve variable $x_i$.

<u>Step 104</u>

The final step, 105, defines the importance of a predictive variable as the product of the slope (Step 1) and the deviance (Step 2) of the variable:

***Importance = Slope * Deviance***

For each policy that is scored, the Importance of each variable is calculated according to the above methodology. The predictive variables are then sorted for every policy according to their Importance measurement to determine which variables contributed the most to the predicted profitability.

Referring to Figs. 3 through 5, the Importance calculation is applied to an exemplary situation illustrating the usage of the proposed Importance calculation in a typical multivariate auto insurance scoring formula. In the example, the following should be assumed: (i) a personal automobile book of business is being analyzed, and (ii) the book has a large quantity of data, e.g., 40,000 data points, available for the analysis. In this example, a linear regression formula is used for its simplicity. As described in more detail below, the scoring formula is given as follows:

$$Y = 0.376 + 0.0061X_1 - 0.0106X_2 + 0.00593X_3 - 0.00334X_4 + 0.011X_5 + 0.075X_6 + 0.049X_7 + 0.027X_8 + 0.0106X_9 + 0.061X_{10} - 0.00242X_{11} - 0.062X_{12} + 0.0109X_{13} + 0.000403X_{14} - 0.00194X_{15} - 0.0017X_{16} + 0.000704X_{17}$$

In the above scoring formula, the target variable, Y, will predict the loss ratio (loss/premium) for a personal automobile policy. A multivariate technique, which can be a traditional linear regression or a more advanced nonlinear technique such as nonlinear regression or neural networks, was used to develop the scoring formula. The formula uses seventeen (17) driver and vehicle characteristics to predict the loss ratio, which are described in Fig. 3.

Any assumptions made for the variables are specified in Fig. 4. For each variable, the information gives a further description of the possible values for each variable based on the total population of the data points used in the model development and stored in database 20. Additionally, Fig. 4 specifies the Mean of the modeling data population and Standard Deviation for each variable.

This example illustrates a "bad" (predicted to be unprofitable) policy having the values for the particular variables specified in Fig. 5. The scoring formula contains a constant term, 0.376, and a parameter for each predictive variable. When the parameter is positive, it indicates that the higher the variable, the higher the Y and hence the worse the predicted profitability. When the parameter is negative, it indicates the opposite. For example, the parameter for vehicle age, $X_2$, is –0.0106. This suggests that the older the vehicle, the lower the Y and the better the profitability. It also suggests that as the vehicle age increases by 1 year, the Y will decrease by 0.0106. On the other hand, the parameter for the number of minor traffic violation, $X_5$, is 0.011. This suggests that the more the violations, the higher the Y and the worse the profitability. It also suggests that as the number of the violation increases by one, the Y will increase by 0.011.

Referring to Fig. 5, the solution of the model indicates that the policy has a predicted loss ratio score of 1.19, which is more than twice the population average of 0.54. A

close review of the seventeen (17) predictive variables further indicates that it has many bad characteristics. For example, it has a number of accidents and violations ($X_5$, $X_6$, $X_9$). It also has a very high number of safety surcharge points ($X_4$) as well as a bad financial credit score ($X_{14}$). Also, the vehicle is very expensive ($X_1$) and the driver is relatively young ($X_{11}$).

While the policy is obviously a bad policy, the unanswered question is which of the seventeen (17) variables are the key driving factors for the bad score? Are the ten (10) driver safety points the number one reason, or the three (3) major violations the number one reason for such a bad score? In addition, what are the top 5 most important reasons? In order to address these questions, the Importance of each variable is calculated using the method described above and in Fig. 2. The first step (102) is to calculate the slope of each predictive variable:

$$\text{Slope of Predictive Variable } x_i = \left| avg\left(\frac{\partial F(X)}{\partial x_i}\right) \right|$$

Since the scoring formula used in the example is a linear formula, the slope is the same as the parameter or coefficient preceding each variable in the scoring formula, as illustrated in column 3 of Fig. 5. The next step (103) is to calculate the deviance for each predictive variable:

$$\text{Deviance of } x_i = \frac{(x_i - \mu_i)}{\sigma_i}$$

where $\mu_i$ is the mean for $x_i$ and $\sigma_i$ is the standard deviation for predicitve variable $x_i$.

The value ($X_i$) for each variable for the sample policy is given in the second column, and the population mean and the population standard deviation are given in columns 3 and 4 of Fig. 4. The calculated slope and deviance for each variable are shown in columns 3 and 4, respectively, of Fig. 5. The next step (104) is to calculate the Importance, which is

the product of slope and deviance. The calculated importance is given in column 5 of Fig. 5. In a final step (105), from the calculated value of the Importance, the variables can be ranked from highest to lowest value as shown in column 6 of Fig. 5.

The ranking is directly translated into a reasons ranking. From column 6, it can be see that the most important reason why the sample policy is a "bad" policy is because the policy has three major traffic ($X_{10}$) violations, compared to the average 0.11 violations for the general population. The second most important reason is that the policy has two no-fault incidences ($X_6$), while the general population on average only has 0.1 violations.

When these two variables are compared to the other fifteen (15) variables, it becomes clear that this policy has values for these two variables that are very different from the general population, as indicated by the high value of deviance. In addition, the parameters (the slopes) for these two variables are also very high, indicating that both variables have a significant impact on the predicted loss ratio and profitability of the policy. In the case of these two variables, the high values of both the slope and the deviance causes these two variables to emerge as the top two most Important factors to explain the bad score for the policy.

With the foregoing method and system an easy–to-understand explanation of which variables are most significant to the score (i.e., Importance) is made available to non-technical end users. Such clear communication and interpretation of insurance profitability scores is critical if they are used by the various interested insurance parties including policyholders, agents, underwriters, and regulators.

***

In so far as embodiments of the invention described herein may be implemented, at least in part, using software controlled programmable processing devices, such as a computer system, it will be appreciated that one or more computer programs for configuring such programmable devices or system of devices to implement the foregoing described methods are to be considered an aspect of the present invention. The computer programs may be embodied as source code and undergo compilation for implementation on processing devices or a system of devices, or may be embodied as object code, for example. Those of ordinary skill will readily understand that the term computer in its most general sense encompasses programmable devices such as those referred to above, and data processing apparatus, computer systems and the like.

Preferably, the computer programs are stored on carrier media in machine or device readable form, for example in solid-state memory or magnetic memory such as disk or tape, and processing devices utilize the programs or parts thereof to configure themselves for operation. The computer programs may be supplied from remote sources embodied in communications media, such as electronic signals, radio frequency carrier waves, optical carrier waves and the like. Such carrier media are also contemplated as aspects of the present invention.

It will thus be seen that the objects set forth above, among those made apparent from the preceding description, are efficiently attained and, since certain changes may be made in carrying out the above method and in the system set forth without departing from the spirit and scope of the invention, it is intended that all matter contained in the above description and shown in the accompanying drawings shall be interpreted as illustrative and not in a limiting sense.

It is also to be understood that the following claims are intended to cover all of the generic and specific features of the invention herein described and all statements of the scope of the invention which, as a matter of language, might be said to fall therebetween.